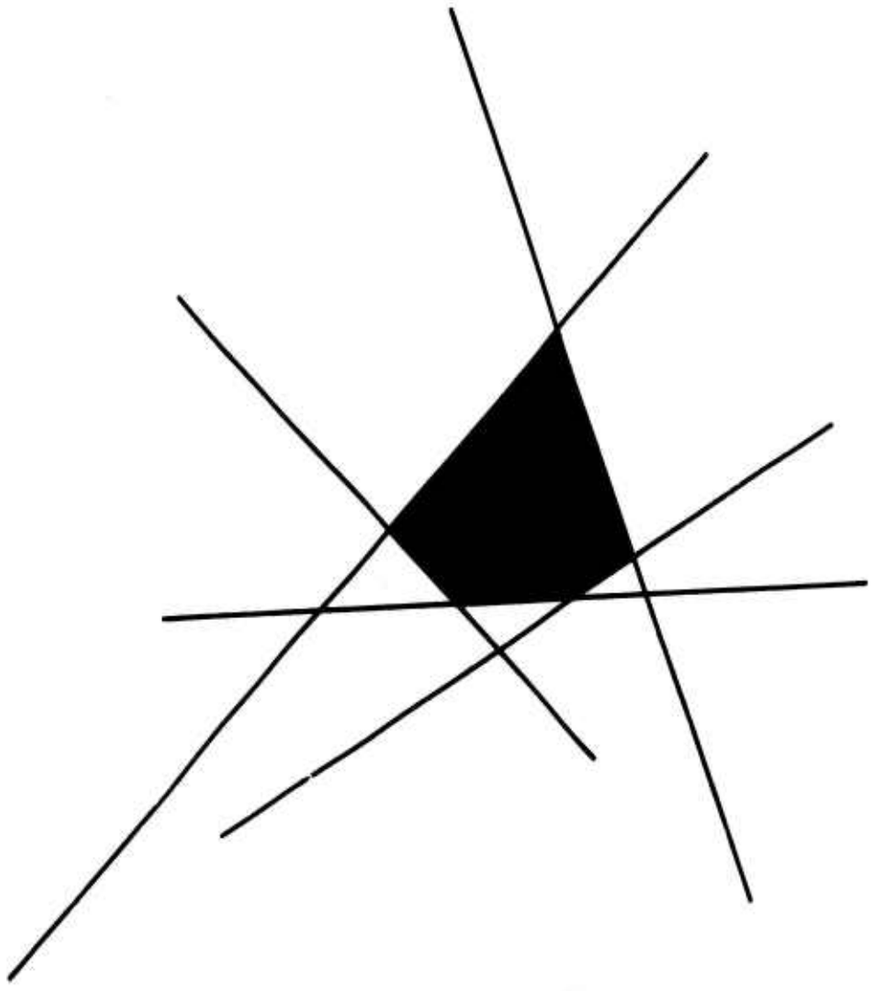
ORC 71-4
APRIL 1971

SOME RESULTS IN DYNAMIC PROGRAMMING

by
SHELDON M. ROSS



OPERATIONS RESEARCH CENTER

COLLEGE OF ENGINEERING
UNIVERSITY OF CALIFORNIA · BERKELEY

19

SOME RESULTS IN DYNAMIC PROGRAMMING

by

Sheldon M. Ross
Department of Industrial Engineering
and Operations Research
University of California, Berkeley

APRIL 1971 ORC 71-4

This research has been partially supported by the Office of Naval Research under Contract N00014-69-A-0200-1010 and the U.S. Army Research Office-Durham under Contract DA-31-124-ARO-D-331 with the University of California. Reproduction in whole or in part is permitted for any purpose of the United States Government.

Unclassified

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)	2a. REPORT SECURITY CLASSIFICATION
University of California, Berkeley	Unclassified
	2b. GROUP

3. REPORT TITLE

SOME RESULTS IN DYNAMIC PROGRAMMING

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

Research Report

5. AUTHOR(S) *(First name, middle initial, last name)*

Sheldon M. Ross

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
April 1971	13	9

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
N00014-69-A-0200-1010	ORC 71-4
b. PROJECT NO. NR 047 033	
c. Research Project No.: RR 003 07 01	9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*
d.	

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
Also supported by the U.S. Army Research Office-Durham under Contract DA-31-124-ARO-D-331.	Mathematical Sciences Division Chief of Naval Research Arlington, Virginia 22217

13. ABSTRACT

SEE ABSTRACT.

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6811

Unclassified

Security Classification

A-31408

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Dynamic Programming						
Positive Dynamic Programming						
Stationary						
Optimal						
Average Cost Optimal						
ε-Optimal						

ABSTRACT

In the first part of this report we consider a dynamic programming model in which all rewards obtained by the decision maker are assumed nonnegative. The decision maker's objective is to successively choose actions so as to maximize his expected reward earned over an infinite time span. It follows from known results that the decision maker's choice need only depend upon the outcome of a randomization that depends on the model only through the state of the model and the time when the choice is made. We show by counterexample that this is basically the smallest class of decision rules that need be considered. Conditions under which a stationary policy is optimal are also presented.

In the second part we consider the same model under a new criteria, namely, the average cost incurred per unit time. An example is presented in which there does not exist an ε-optimal randomized stationary policy.

TABLE OF CONTENTS

SOME REMARKS ON POSITIVE DYNAMIC PROGRAMMING

by

Sheldon M. Ross

1. INTRODUCTION

Consider a Markov decision process having a countable state space I and a finite action space A. If action a is chosen when in state i, then

(i) we receive a nonnegative bounded reward $R(i,a)$

and

(ii) the next state of the system is chosen according to the transition probabilities $\{P_{ij}(a), j \in I\}$.

A policy is any measurable rule for choosing actions. Let X_t and a_t denote, respectively, the state of the process and the action chosen at time t. For any policy π, we define

$$V_\pi(i) = E_\pi\left[\sum_{t=0}^{\infty} R(X_t,a_t) \mid X_0 = i\right]. \tag{1}$$

A policy is said to be

(1) *stationary*, if the action it chooses at any time is a deterministic function of the state at that time.

(2) *randomized stationary*, if its action at any time is a randomized function of the state at that time.

(3) *Markov* or *memoryless*, if its action at time t is a deterministic function of the state at time t and t.

(4) *randomized Markov* or *randomized memoryless*, if its action at time t is a randomized function of the state at time t and t.

It follows from results presented by Derman and Strauch [3] that we need never go outside the class of randomized memoryless policies. That is, for any policy π, there exists a randomized memoryless policy π' such that

$$V_{\pi'}(i) \geq V_{\pi}(i) \qquad \forall\, i\,. \tag{2}$$

They prove this by showing that the class of randomized memoryless policies is large enough so that, for any policy π, there exists a randomized memoryless policy π' such that

$$P_{\pi}\{X_t = j, a_t = a \mid X_0 = i\} = P_{\pi'}\{X_t = j, a_t = a \mid X_0 = i\}$$

which, of course, implies that $V_{\pi}(i) = V_{\pi'}(i)$.

In Section 2 we show, by counterexample, that we cannot generally restrict attention to either the class of randomized stationary policies or to the class of memoryless policies. In Section 3 we prove some results concerning the existence of a stationary policy that maximizes (1).

2. THE COUNTEREXAMPLES

The first counterexample shows that we cannot always restrict attention to the memoryless policies.

Example 1:

Let the states be given by $0,1,1',2,2', \ldots$. State 0 is an absorbing state and once entered can never be left, i.e.,

$$P_{00} = 1 .$$

In state $n, n > 0$, there are 2 possible actions having respective transition probabilities

$$P_{n,n+1}(1) = P_{n,n'}(2) = 1 , \qquad n > 0 .$$

In state $n', n > 0$, there is a single available action, having transition probabilities

$$P_{n',(n-1)'} = 1 \qquad n > 1$$

$$P_{1',0} = 1 \qquad n = 1 .$$

The rewards depend only on the state and are given by

$$R(n) = 0 \qquad n \geq 0$$

$$R(n') = 1 \qquad n > 0 .$$

Suppose the initial state is state 1 . It is easy to see that under any memoryless rule the total expected reward will be finite. However the randomized stationary policy which, when in state n , selects action 1 with probability a_n and action 2 with probability $1 - a_n$ has an infinite expected return when the a_n are chosen so that

$$\prod_{i=1}^{n} a_i \to 0 \quad \text{as} \quad n \to \infty$$

and

$$\sum_{n=1}^{\infty} \prod_{i=1}^{n} a_i = \infty .$$

The second example shows that we cannot always restrict attention to the randomized stationary policies.

Example 2:

The states are given by $1,2,3, \ldots, \infty$. In state n there are 2 possible actions having respective transition probabilities

$$P_{n,n+1}(1) = 1 \qquad 1 \leq n < \infty$$

$$P_{n,1}(2) = \alpha_n = 1 - P_{n,\infty}(2) \qquad 1 \leq n < \infty .$$

State ∞ is an absorbing state, i.e.,

$$P_{\infty\infty} = 1 .$$

The rewards depend only on the state and are given by

$$R(1) = 1$$

$$R(n) = 0 \qquad n = 2,3, \ldots, \infty .$$

The values α_n are chosen to satisfy

$$\prod_{n=1}^{\infty} \alpha_n > 0 , \qquad \alpha_n < 1 \text{ all } n . \tag{3}$$

Suppose that the initial state is state 1 . It is easy to see that under any randomized stationary policy the expected number of visits to state 1 is a geometric random variable with finite means, hence the total expected return is finite. However, consider the policy which on its nth return to state 1 chooses action 1 n times and then chooses action 2 . Since, by (3) this policy has a positive probability of visiting state 1 infinitely often, it has an infinite expected return.

3. CONDITIONS

Let

$$V(i) = \sup_{\pi} V_{\pi}(i) , \qquad i \in I .$$

It has been shown by Blackwell [2] that V is the smallest nonnegative solution of

$$V(i) = \max_{a} \left\{ R(i,a) + \sum_{j} P_{ij}(a) V(j) \right\} , \qquad i \in I . \tag{4}$$

Let π^* be a stationary policy which, when in state i, selects an action maximizing the right side of (4).

Proposition 1:

If $V(i) < \infty$, then $V_{\pi^*}(i) = V(i)$ if and only if

$$E_{\pi^*}[V(X_n) \mid X_0 = i] \to 0 \qquad \text{as} \quad n \to \infty .$$

Proof:

Let $H_t = \{X_0, a_0, X_1, a_1, \ldots, X_t, a_t\}$ denote the history of the process up to time t. Now

$$E_{\pi^*} \sum_{t=1}^{n} \left[V(X_t) - E_{\pi^*}(V(X_t) \mid H_{t-1}) \right] = 0 .$$

However,

$$\begin{aligned} E_{\pi^*}[V(X_t) \mid H_{t-1}] &= \sum_{j} P_{X_{t-1} j}(a_{t-1}) V(j) + R(X_{t-1}, a_{t-1}) - R(X_{t-1}, a_{t-1}) \\ &= V(X_{t-1}) - R(X_{t-1}, a_{t-1}) . \end{aligned}$$

Thus

$$E_{\pi^*}\sum_{t=1}^{n} R(X_{t-1},a_{t-1}) = E_{\pi^*}[V(X_0)] - E_{\pi^*}[V(X_n)]$$

and the result follows by letting $n \to \infty$.

Corollary 1:

Let $W(i), i \in I$, be any finite nonnegative solution of

$$W(i) = \max_a \left\{R(i,a) + \sum_j P_{ij}(a)W(j)\right\} , \qquad i \in I \tag{5}$$

and let π^* be a stationary policy which, when in state i , selects an action maximizing the right side of (5). If

$$E_{\pi^*}[W(X_n) \mid X_0 = i] \to 0 \qquad \text{as} \quad n \to \infty$$

then

$$V_{\pi^*}(i) = V(i) = W(i) .$$

Proof:

By the same reasoning as in Proposition 1, we obtain that

$$E_{\pi^*}\sum_{t=1}^{n} R(X_{t-1},a_{t-1}) = E_{\pi^*}[W(X_0)] - E_{\pi^*}[W(X_n)] .$$

The result now follows by letting $n \to \infty$ and recailing that V is the smallest nonnegative solution of (5).

Corollary 2:

Suppose there exists a (stopped) state -- call it 0 -- which is such that

$$P_{00}(a) = 1 \qquad \text{all } a \in A$$

$$R(0,a) = 0 \qquad \text{all } a \in A .$$

Let π^* be a stationary policy determined by the optimality Equation (4). Then if

(i) V is bounded

(ii) $P_{\pi^*}[\lim X_n = 0 \mid X_0 = i] = 1$

then

$$V_{\pi^*}(i) = V(i) .$$

Proof:

The proof follows from Proposition 1 and the bounded convergence theorem since $V(0) = 0$.

REFERENCES

[1] Blackwell, David, "Positive Dynamic Programming," PROCEEDINGS OF THE FIFTH BERKELEY SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY, Vol. 1, pp. 415-418, University of California Press, (1967).

[2] Blackwell, David, "On Stationary Strategies," Royal Statistical Society Journal, Series A, (1970).

[3] Derman, C. and R. Strauch, "A Note on Memoryless Rules for Controlling Sequential Control Processes," Annals of Mathematical Statistics, Vol. 37, pp. 276-279, (1966).

[4] Ornstein, Donald, "On the Existence of Stationary Optimal Strategies," Proceedings of the American Mathematical Society, Vol. 20, pp. 563-569, (1969).

[5] Strauch, Ralph, "Negative Dynamic Programming," Annals of Mathematical Statistics, Vol. 37, pp. 871-889, (1966).

ON THE NONEXISTENCE OF ε-OPTIMAL RANDOMIZED STATIONARY POLICIES IN AVERAGE COST MARKOV DECISION MODELS

by

Sheldon M. Ross

1. INTRODUCTION

Consider a Markov decision process [see Derman (1966) or Ross (1968)] having a countable state space I and a finite action space A. If action a is chosen when in state i, then

(i) a cost $c[i,a]$ is incurred, and

(ii) the next state is determined according to the transition probabilities $\{P_{ij}(a), j \in I\}$.

A policy is any measurable rule for choosing actions, and is called stationary if the (possibly randomized) action the policy chooses at any time depends only on the state of the process at that time. In Maitra (1966), the question is asked of whether or not there always exists an ε-optimal stationary policy under the average expected cost criterion. That is, is there a stationary policy whose average expected cost is within ε of the infimum over all policies? We answer this in the negative by the following counterexample.

2. THE COUNTEREXAMPLE

Let the states be given by $1,1',2,2', \ldots, n,n', \ldots, \infty$. In state n, $1 \leq n < \infty$, there are two actions, with transition probabilities given by

$$P_{n,n+1}(1) = 1$$

$$P_{n,n'}(2) = \alpha_n = 1 - P_{n,\infty}(2) .$$

In state n', there is a single action, having transition probabilities

$$P_{n',(n-1)'} = 1 \ , \ n \geq 2$$

$$P_{1',1} = 1 \ .$$

State ∞ is an absorbing state and once entered is never left, i.e.,

$$P_{\infty\infty} = 1 \ .$$

The costs depend only on the state and are given by

$$c[n,a] = 2 \quad \text{all} \quad n = 1,2, \ldots, \infty \ , \text{ all actions } a$$

$$c[n',a] = 0 \quad \text{all} \quad n \geq 1 \ , \text{ all } a \ .$$

The values α_n are chosen to satisfy

(i) $\alpha_n < 1$

(ii) $\prod_{n=1}^{\infty} \alpha_n = 3/4 \ .$

Suppose the initial state is state 1. If a stationary policy is employed then, with probability 1, a cost of 2 will be incurred in all but a finite number of time periods. This follows, since under a stationary policy, each time the process enters state 1 there is a fixed positive probability that the process will never again re-enter that state. Therefore, under a stationary policy, the average cost will, with probability 1, equal 2. Hence, by the bounded convergence theorem, the average expected cost will also equal 2.

Now let R be the nonstationary policy which initially chooses action 2, and then on its nth return to state 1, chooses action 1 n times and then chooses action 2. The average cost under this policy will equal

$$\begin{cases} 2 & \text{with probability } 1 - \prod_{n=1}^{\infty} \alpha_n \\ 1 & \text{with probability } \prod_{n=1}^{\infty} \alpha_n . \end{cases}$$

This is true since $\prod_{n=1}^{\infty} \alpha_n$ represents the probability that, under R, the process will never enter state ∞. Hence, by the bounded convergence theorem, the average expected cost under R is $3/4 + 2/4 = 5/4$.

Hence, there is no ε-optimal randomized stationary policy for $\varepsilon < 3/4$.

REFERENCES

[1] Derman, Cyrus, "Denumerable State Markovian Decision Processes-Average Cost Criterion," Ann.Math.Statist., Vol. 37, pp. 1545-1556, (1966).

[2] Fisher, L. and S. Ross, "An Example in Denumerable Decision Processes," Ann.Math.Statist., Vol. 39, pp. 674-675, (1968).

[3] Maitra, Ashok, "A Note on Undiscounted Dynamic Programming," Ann.Math.Statist., Vol. 37, pp. 1042-1044, (1966).

[4] Ross, Sheldon, "Nondiscounted Denumerable Markovian Decision Models," Ann.Math.Statist., Vol. 39, pp. 412-423, (1968).